Efficient Optimization of Constrained Nonlinear Resource Allocations *

Mung Chiang
Electrical Engineering Department, Princeton University, NJ 08544

Abstract

We present an efficient method to optimize network resource allocations under nonlinear Quality of Service (QoS) constraints. We first propose a suite of generalized proportional allocation schemes that can be obtained by minimizing the information-theoretic function of relative entropy. We then optimize over the allocation parameters, which are usually design variables an engineer can directly vary, either for a particular user or for the worst-case user, under constraints that lower bound the allocated resources for all other users. Despite the nonlinearity in the objective and constraints, we show this suite of resource allocation optimization can be efficiently solved for global optimality through a convex optimization technique called geometric programming.

This general method and its extensions are applicable to a wide array of resource allocation problems, including processor sharing, congestion control, admission control, and wireless network power control. We focus on several specific formulations and numerical examples for an admission control scheme, and for power control problems of throughput maximization under outage and delay constraints for wireless multihop networks.

Keywords: Connection admission control, Convex optimization, Geometric programming, Power control, Proportional fairness, Resource allocation, Wireless ad hoc networks.

*Parts of this work have appeared in the Proceedings of *IEEE INFOCOM* New York July 2002 and *IEEE GLOBECOM* San Francisco December 2003.

1 Introduction

1.1 Overview

Consider n users indexed by i sharing a common pool of communication network resource X, such as bandwidth or buffer. The amount of resource allocated to connection i is denoted by x_i. Based on several motivating examples in subsection 2.1, we will introduce the following Generalized Proportional Allocation (GPA) form, where the total resource X is allocated to connection i in proportion to some p_i and normalized by a sum of parameters $\sum_j \gamma_{ji} + \alpha\nu_i$:

$$x_i = \frac{p_i}{\sum_j \gamma_{ji} + \alpha\nu_i} X,$$

where the allocation parameters $p_i, \nu_i, \gamma_{ji} \geq 0$ belong to fixed ranges of values for each user i (different ranges for different QoS classes), and $\alpha \geq 0$ is a given weight. This form of resource allocation appears in many applications, and we show in subsection 2.2 that we will obtain a GPA form whenever we solve an underlying optimization problem that minimizes relative entropy, a fundamental function in information theory [5].

These allocation parameters p_i, ν_i, γ_{ji} can be further optimized to maximize the resource received by a particular user i^* in the highest Quality of Service (QoS) class, subject to constraints that lower bound the resources received by each of the other users. Alternatively, for maxmin fairness, allocation parameters can be optimized to maximize the resource received by the user with the minimum received resource. Although both versions are nonlinear problems, which in general could be difficult to solve and could take exponential-time algorithms to find a global optimality, subsection 2.3 uses geometric programming to show that they have the following desirable properties:

- Every locally optimal allocation is also globally optimal, which can be efficiently obtained in polynomial time through convex optimization techniques.
- Bottlenecks of resource allocation constraints are readily detected, so that if additional resources becomes available, we know where to allocate them to alleviate the bottlenecks of resource demands.

We apply this general method of resource allocation first to an admission control scheme in subsection 3.1 to balance the total admitted rate and fairness among the competing connections, and then to power control for wireless multihop networks in subsections 3.2 and 3.3 to maximize system throughput under channel outage and average delay constraints.

We will use the following notation. Given two column vectors $\mathbf{x}$ and $\mathbf{y}$ of length n, we express the sum $\sum_{i=1}^n x_i y_i$ as an inner product $\mathbf{x}^T\mathbf{y}$. Componentwise inequalities on a vector $\mathbf{x}$ with n entries is expressed using the $\succeq$ symbol: $\mathbf{x} \succeq 0$ denotes $x_i \geq 0, i = 1, 2, \ldots, n$.

1.2 Geometric programming

Both the theoretical results and numerical algorithms in this paper use the tools of Lagrange duality, convex optimization [1], and geometric programming [6]. First recall that minimizing a convex objective function subject to upper bounds on convex constraint functions can be

easy. It is easy in theory because a local minimum is a global minimum. If the objective function is strictly convex, there is a unique globally optimal solution to the nonlinear problem. A convex optimization problem can also be easy to solve in practice, when put in the right form with the right input data structure, because then there are fast algorithms, such as the primal-dual interior point method [13], that find a globally optimal solution in polynomial time. More importantly, empirical evidence shows that the running times of such algorithms grow very slowly with the problem sizes. Appropriately formulated convex optimization problems are intrinsically tractable and can be efficiently solved. In addition, there is a useful Lagrange duality theory for convex optimization, where we can solve an optimization problem through its Lagrange dual problem. In this paper, we extensively use a special type of nonlinear optimization called geometric programming [1, 6] which has recently found several applications in communication systems [2, 3, 4, 8, 9].

We first define a monomial as a function $f : \mathbf{R}_+^n \to \mathbf{R}$:

$$f(\mathbf{x}) = dx_1^{a^{(1)}} x_2^{a^{(2)}} \dots x_n^{a^{(n)}},$$

where $d \geq 0$ and $a^{(j)} \in \mathbf{R}$. A posynomial is a sum of monomials

$$f(\mathbf{x}) = \sum_{k=1}^{K} d_k x_1^{a_k^{(1)}} x_2^{a_k^{(2)}} \dots x_n^{a_k^{(n)}},$$

where $d_k \geq 0,\ k = 1, 2, \dots, K$, and $a_k^{(j)} \in \mathbf{R},\ j = 1, 2, \dots, n, k = 1, 2, \dots, K$.

Geometric programming in standard form is an optimization problem in the following form:

$$\begin{array}{ll} \text{minimize} & f_0(\mathbf{x}) \\ \text{subject to} & f_i(\mathbf{x}) \leq 1, \quad i = 1, 2, \dots, M_1, \\ & h_l(\mathbf{x}) = 1, \quad l = 1, 2, \dots, M_2 \end{array} \tag{1}$$

where f_0 and f_i are posynomials and h_l are monomials in variables $\mathbf{x}$.

Geometric programming in the above standard form is not a convex optimization problem. However, with a logarithmic change of the variables: $y_i = \log x_i, b_{ik} = \log d_{ik}$, we can turn it into the following convex form:

$$\begin{array}{ll} \text{minimize} & p_0(\mathbf{y}) = \log \sum_k \exp(\mathbf{a}_{0k}^T \mathbf{y} + b_{0k}) \\ \text{subject to} & p_i(\mathbf{y}) = \log \sum_k \exp(\mathbf{a}_{ik}^T \mathbf{y} + b_{ik}) \leq 0, \quad i = 1, 2, \dots, M_1, \\ & q_l(\mathbf{y}) = \mathbf{a}_l^T \mathbf{y} + b_l = 0, \quad l = 1, 2, \dots, M_2 \end{array} \tag{2}$$

where the optimization variables are $\mathbf{y}$. The logarithm of a sum of exponentials is a convex function [1, 2, 6], thus (2) is a convex optimization problem. Many nice properties of geometric programs are also maintained for an extended geometric program [6], where the objective function can be a posynomial plus the logarithm of a monomial.

2 Generic Framework

2.1 Generalized proportional allocation

We first review several examples of resource allocation that have found wide use in communication systems.

Example 1: Recall the Generalized Processor Sharing scheme [14] where an egress link with a total rate of R is shared among multiple connections each receiving rate R_i:

$$R_i = \frac{\phi_i}{\sum_j \phi_j} R$$

where $\{\phi_i\}$ are the parameters that can be directly varied in the system design to produce a desirable $\{R_i\}$. Rates are allocated proportional to ϕ_i and then normalized.

Example 2: [12] shows that rate control on a single link through the protocol of TCP Reno produces the following rate allocation proportional to the inverse of round trip delay D_i:

$$R_i = \frac{\frac{1}{D_i}}{\sum_j \frac{1}{D_j}} R.$$

Example 3: A slightly more complicated form of proportional allocation occurs in wireless networks, where the received 'resource' for the user on link i is the Signal to Interference Ratio:

$$\text{SIR}_i = \frac{P_i G_{ii}}{\sum_{j \neq i} P_j G_{ij} + n_i},$$

P_i is the transmit power and n_i the noise on link i, and G_{ij} is the path gain from the sender on link j to the receiver on link i. Again, transmit powers $\{P_i\}$ are the parameters that can be directly varied in the system design to produce a desirable $\{\text{SIR}_i\}$, which in turn determine other QoS metrics such as attainable data rates.

In this section, we first give a general parametrization of resource allocation that includes these and other examples as special cases, then determine what problem would lead to an allocation in this form, and finally optimize over the allocation parameters for efficient resource allocation. We first propose the following

Definition 1 *Given $p_i, \gamma_{ji}, \nu_i \geq 0, \forall i, j$, and $\alpha \geq 0$, the Generalized Proportional Allocation (GPA) form of allocating total resource X into resource x_i for each user i is*

$$x_i = \frac{p_i}{\sum_j \gamma_{ji} + \alpha \nu_i} X. \tag{3}$$

Obviously, the processor sharing example follows the GPA form with $p_i = \phi_i, \gamma_{ji} = \phi_j, \forall i, \alpha = 0$ and $X = R$. The TCP Reno example follows the GPA form with $p_i = \frac{1}{D_i}, \gamma_{ji} = \frac{1}{D_j}, \forall i$, $\alpha = 0$, and $X = R$. The wireless SIR example follows the GPA form with $p_i = P_i, \gamma_{ji} = P_j \frac{G_{ij}}{G_{ii}}, j \neq i, \gamma_{ii} = 0, \alpha = 1, \nu_i = \frac{N_i}{G_{ii}}$, and $X = 1$. In each of these examples, at least a subset of the allocation parameters are the variables directly adjustable by the system designer to induce a desirable resource allocation $\{x_i\}$, which are not directly controllable themselves.

2.2 A relative entropy minimization

Since the GPA form (3) includes various examples as special cases, we wonder what type of problems leads to solutions in this form. We will show that solving the following optimization leads to a resource allocation in the GPA form.

Definition 2 *Relative Entropy Minimization (REM) is a convex optimization problem in the following form:*

$$\begin{array}{ll} \textit{minimize} & D(\mathbf{p}\|\mathbf{x}) + \alpha(\boldsymbol{\nu}^T\mathbf{x}) \\ \textit{subject to} & \mathbf{A}\mathbf{x} \preceq \mathbf{v}, \quad \mathbf{x} \succeq 0 \end{array} \tag{4}$$

where the optimization variables are $\mathbf{x}$*, and the constant parameters are* $\mathbf{p}, \mathbf{A}, \mathbf{v}, \boldsymbol{\nu}, \alpha \succeq 0$*. Relative entropy is defined for* $\mathbf{p}, \mathbf{q} \succeq 0$ *as*

$$D(\mathbf{p}\|\mathbf{q}) = \sum_i p_i \log \frac{p_i}{q_i}.$$

As can easily verified, REM can also be written as an extended geometric program of minimizing a posynomial and the log of a monomial in variables $\mathbf{x}$:

$$\begin{array}{ll} \text{minimize} & \alpha(\boldsymbol{\nu}^T\mathbf{x}) + \sum_i p_i \log p_i + \log \prod_i x_i^{-p_i} \\ \text{subject to} & \mathbf{A}\mathbf{x} \preceq \mathbf{v}, \quad \mathbf{x} \succeq 0. \end{array} \tag{5}$$

Proposition 1 *A resource allocation in the GPA form (3) is obtained if a REM problem (4) is solved.*

Proof. We first show that the Lagrange dual problem of (4) is

$$\begin{array}{ll} \text{maximize} & \sum_i p_i \log \beta_i - \mathbf{v}^T\boldsymbol{\lambda} \\ \text{subject to} & \mathbf{A}^T\boldsymbol{\lambda} + \alpha\boldsymbol{\nu} = \boldsymbol{\beta}, \\ & \boldsymbol{\lambda} \succeq 0 \end{array} \tag{6}$$

where the optimization variables are $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$, and the constant parameters are $\mathbf{A}, \mathbf{v}, \mathbf{p}, \boldsymbol{\nu}$ and α. Furthermore, the optimal primal variables $\mathbf{x}^*$ can be obtained from the optimal dual variables $\boldsymbol{\lambda}^*$ as

$$x_i^* = \frac{p_i}{\sum_j \lambda_j^* A_{ji} + \alpha\nu_i}.$$

Indeed, let us form the Lagrangian of the primal problem (4), ignoring the constant term of $\sum_i p_i \log p_i$:

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\sigma}) = -\sum_i p_i \log x_i + \alpha(\boldsymbol{\nu}^T\mathbf{x}) + \boldsymbol{\lambda}^T(\mathbf{A}\mathbf{x} - \mathbf{v}) - \boldsymbol{\sigma}^T\mathbf{x}$$

where $\boldsymbol{\lambda}, \boldsymbol{\sigma} \succeq 0$ are the Lagrange multiplier vectors. Let the derivative of $L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\sigma})$ with respect to x_i be equal to 0, we obtain

$$x_i = \frac{p_i}{\sum_j \lambda_j A_{ji} + \alpha\nu_i - \sigma_i}.$$

Substitute this $\mathbf{x}$ into the Lagrangian $L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\sigma})$, we obtain the Lagrange dual function $g(\boldsymbol{\lambda}, \boldsymbol{\sigma})$:

$$-\sum_i p_i \log \frac{p_i}{\sum_j \lambda_i A_{ji} + \alpha\nu_i - \sigma_i} + \sum_j \lambda_j \sum_i A_{ji} \frac{p_i}{\sum_j \lambda_j A_{ji} + \alpha\nu_i - \sigma_i} - \boldsymbol{\lambda}^T\mathbf{v} - \sum_i \sigma_i \frac{p_i}{\sum_j \lambda_i A_{ji} + \alpha\nu_i - \sigma_i},$$

which can be simplified to

$$g(\boldsymbol{\lambda}, \boldsymbol{\sigma}) = \sum_i p_i \log \left(\sum_j \lambda_j A_{ji} + \alpha\nu_i - \sigma_i \right) - \mathbf{v}^T\boldsymbol{\lambda} + \sum_i p_i - \sum_i p_i \log p_i.$$

Therefore, the Lagrange dual problem can be stated as

$$\begin{array}{ll} \text{maximize} & \sum_i p_i \log\left(\sum_j \lambda_j A_{ji} + \alpha\nu_i - \sigma_i\right) - \mathbf{v}^T\boldsymbol{\lambda} \\ \text{subject to} & \boldsymbol{\sigma}, \boldsymbol{\lambda} \succeq 0. \end{array}$$

Since the objective function is a non-increasing function of $\boldsymbol{\sigma} \succeq 0$, we let $\boldsymbol{\sigma} = 0$, and simplify the Lagrange dual problem to

$$\begin{array}{ll} \text{maximize} & \sum_i p_i \log\left(\sum_j \lambda_j A_{ji} + \alpha\nu_i\right) - \mathbf{v}^T\boldsymbol{\lambda} \\ \text{subject to} & \boldsymbol{\lambda} \succeq 0. \end{array}$$

Now letting $\boldsymbol{\beta} = \mathbf{A}^T\boldsymbol{\lambda} + \alpha\boldsymbol{\nu}$ proves the claim.

Therefore, for every REM problem (4), there corresponds a proportional allocation in the parameterized form of (3). In particular, the network utility model in Kelly [10] is a special case of REM where $\alpha = 0$ and $\mathbf{A}$ is a $0-1$ matrix denoting the routing decisions.

2.3 Efficient optimization of constrained resource allocation

Suppose we have solved a REM problem (4) and obtained a resource allocation in the GPA form (3). A desirable next step is to vary the allocation parameters to optimize for the performance of a 'premium class' user, subject to the QoS constraints of minimum resource requirements $x_{i,min}$ for all other users i, and the range constraints on the allocation parameters themselves. Alternatively, for maxmin fairness, the objective is to maximize the resource allocated to the user with the minimal received resource, whichever turns out to be the worst-case user. Usually, the constraints of resource demands from different users compete against each other (*i.e.*, γ_{ji} are increasing functions of p_j). Note that the constant parameters $\mathbf{p}, \mathbf{A}$ and $\boldsymbol{\nu}$ in REM (4) now become the optimization variables. Indeed, in many instances of the GPA form of resource allocation, at least a subset of the allocation parameters (*e.g.*, transmit powers in Example 3) are design variables that can be directly controlled by an engineer. Obviously, the allowed range of allocation parameters for a higher QoS class user will be higher.

Specifically, assuming that $\{x_i\}$ are allocated according to the GPA form, and that user i^* is the highest QoS class user, we have the following **generic problem of constrained resource allocation**:

$$\begin{array}{ll} \text{maximize} & x_{i^*} \quad (\text{or maximize } \min_i x_i) \\ \text{subject to} & x_i \geq x_{i,min}, \quad \forall i, \\ & \text{variables } (p_i, \gamma_{ji}, \nu_i) \geq \text{lower bounds}, \\ & \text{variables } (p_i, \gamma_{ji}, \nu_i) \leq \text{upper bounds}. \end{array} \tag{7}$$

Because the GPA form is nonlinear (*i.e.*, $\mathbf{x}$ are nonlinear functions of the variables $\mathbf{p}, \boldsymbol{\gamma}, \boldsymbol{\nu}$), the above generic optimization is a suite of nonlinear problems, which in general take exponential-time algorithms to solve for global optimality. For example, it is known that even determining the feasibility of the competing constraints in the case of wireless power control in Example 3 is difficult. However, we show the following

Proposition 2 *The optimization problem (7) for resource allocation in the GPA form (3) is a geometric program, thus can be turned into a convex optimization problem. Therefore, every locally optimal allocation is a globally optimal one, which can be efficiently (and in provably polynomial time) computed through interior point algorithms.*

Proof. The claim is readily verified if the objective in (7) is to maximize x_{i^*}. In this case, omitting the monomial constraints in the form of range constraints on the variables, we can rewrite (7) as

$$\begin{array}{ll} \text{minimize} & \frac{1}{x_{i^*}} \\ \text{subject to} & \frac{1}{x_i} \leq \frac{1}{x_{i,min}} \quad i = 1, 2, \ldots, N. \end{array} \tag{8}$$

By the structure of GPA forms, x_i are inverted posynomials of the variables $\mathbf{p}, \boldsymbol{\nu}$ and γ_{ij}, thus (8) is minimizing a posynomial subject to upper bound constraints on other posynomials. Therefore, (7) is equivalent to a geometric program, which can in turn be converted into a convex optimization problem.

To prove the claim for the maxmin fairness case, we can use the following technique to convert the problem of maximizing (over variables $\mathbf{z}$) the minimum of $g_j(\mathbf{z})$ to be maximizing over $(\mathbf{z}, t)$ (where t is an auxiliary variable) such that $g_j(\mathbf{z}) \geq t, \forall j$. Specifically, for the maxmin fair optimization, the following problem

$$\begin{array}{ll} \text{maximize} & \min_{j=1,2,\ldots,M} g_j(\mathbf{z}) \\ \text{subject to} & f_i(\mathbf{z}) \geq 1, \quad i = 1, 2, \ldots, N \end{array} \tag{9}$$

where the optimization variables are $\mathbf{z}$, and g_j, f_i are inverted posynomials, is easily verified to be equivalent to the following problem:

$$\begin{array}{ll} \text{maximize} & t \\ \text{subject to} & g_j(\mathbf{z}) \geq t, \quad j = 1, 2, \ldots, M, \\ & f_i(\mathbf{z}) \geq 1, \quad i = 1, 2, \ldots, N \end{array} \tag{10}$$

where the optimization variables are $\mathbf{z}$ and t. Now we rewrite the optimization (10) as

$$\begin{array}{ll} \text{minimize} & t^{-1} \\ \text{subject to} & \frac{t}{g_j(\mathbf{z})} \leq 1, \quad j = 1, 2, \ldots, M, \\ & \frac{1}{f_i(\mathbf{z})} \leq 1, \quad i = 1, 2, \ldots, N. \end{array}$$

The objective function is a monomial, and the inequality constraints are posynomials of $(\mathbf{z}, t)$. Therefore, this is a geometric program in standard form.

It is worth noting that optimization (7) may not have any feasible solution. Indeed, if the QoS constraints $x_i \geq x_{i,min}, \forall i$ are too strict, there may exists no resource allocation that simultaneously meets all the constraints. Fortunately, due to the geometric programming nature of the problem, feasibility of resource allocation in the GPA form can also be efficiently determined, and used for admission control and pricing in a network: a new user is admitted into the system only if the resulted new problem (7) is still feasible, and the user is charged in proportion to the resulted reduction in the objective value of (7). Examples of such admission and pricing schemes will be given in subsection 3.2.

A geometric program can also be efficiently solved through its Lagrange dual problem, where we associate a Lagrange dual variable $\sigma_i \geq 0$ for each resource demand constraint $x_i \geq x_{i,min}$. By complementary slackness [1], if an optimal dual variable $\sigma_1^* > 0$, then we know that the QoS requirement constraint for user 1 is tight at optimality, *i.e.*, $x_1^* = x_{1,min}$.

The scope of the above generic method of efficient resource allocation can be extended to accommodate several types of nonlinear functions of the GPA form (which are in turn nonlinear functions of the underlying design variables) in the objective and constraints. This is shown through specific formulations and numerical examples in the next section.

3 Application Examples

3.1 Application 1: Admission control

In this subsection, we show a simple application of the geometric programming method of resource allocation for Connection Admission Control (CAC). Consider the ingress of either a switch or a network as shown in Figure 1. There are K connections trying to get admitted into the system. They first pass through a flow control mechanism, such as leaky buckets, to conform the connections to their respective provisioned rates λ_i specified in the QoS service level agreement. Due to limitation in the available resource, the CAC controller has to enforce admission control among the contending connections.

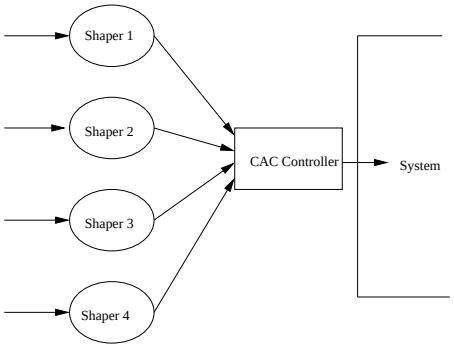

Figure 1: Flows contending for admission into the system through traffic shapers and a connection admission controller.

Consider the following simple CAC algorithm that leads to a rate allocation in the GPA form. The CAC controller has an exponential service time with rate μ. If the first service time of the CAC controller occurs before any packet from the contending connections arrive at the controller, no connection will be admitted to the system. However, if packets from some connections arrive before the first service time of the CAC controller, then the connection whose packet arrives first will be admitted and the other connections will not be admitted.

Lemma 1 *The total rate of admission R_a and the rate of admission for each connection R_i (normalized by the maximum total rate) are of the following GPA forms:*

$$\begin{aligned} R_a &= \frac{\sum_{k=1}^{K} \lambda_k}{\sum_{k=1}^{K} \lambda_k + \mu}, \\ R_i &= \frac{\lambda_i}{\sum_{k=1}^{K} \lambda_k + \mu}. \end{aligned}$$

Intuitively, the relative magnitudes of μ and $\sum_{i=1}^{K} \lambda_i$ determine the admission rate. The relative magnitudes among λ_i determine the fairness among the connections. The parameter μ can be set based on the system congestion condition, and λ_i can be set based on the QoS provisioning terms. Following the geometric programming method in subsection 2.3, we show how the parameters μ and λ_i can be dynamically optimized to provide a flexible control of both the total admission rate and a fair rate allocation among the contending connections.

Formulation 1 *The following nonlinear problem of maximizing the admission rate for a particular connection i^*, subject to the total admission rate constraint $R_{a,max}$, the QoS constraints of guaranteed rate $R_{i,min}$ for each connection i, and the range constraints on variables λ_i and μ, can be efficiently solved for global optimality as a geometric program:*

$$\begin{array}{ll} \textit{maximize} & R_{i^*}(\boldsymbol{\lambda}, \mu) \\ \textit{subject to} & R_a(\boldsymbol{\lambda}, \mu) \leq R_{a,max}, \\ & R_i(\boldsymbol{\lambda}, \mu) \geq R_{i,min}, \quad \forall i, \\ & \lambda_{i,max} \geq \lambda_i \geq \lambda_{i,min}, \quad \forall i, \\ & \mu_{max} \geq \mu \geq \mu_{min}. \end{array} \tag{11}$$

Note that although the first constraint in Formulation 1 is an upper bound instead of the lower bounds on inverted posynomials as in (7), it is still equivalent to a posynomial upper bound $(1 - R_{a,max}) \left(\sum_{j=1}^{K} \lambda_j \mu^{-1} + 1 \right) \leq 1$ due to the specific GPA structure in this case.

Formulation 2 *The following nonlinear problem of maximizing the admission rate for the worst-case connection can be efficiently solved for global optimality as a geometric program:*

$$\begin{array}{ll} \textit{maximize} & \min_i R_i(\boldsymbol{\lambda}, \mu) \\ \textit{subject to} & \textit{Same constraints as in Formulation 1.} \end{array} \tag{12}$$

Note that that the parameters $\lambda_{i,min}, \lambda_{i,min}, \mu_{max}$ and μ_{max} determine the ranges over which λ_i and μ can vary. Larger the $\lambda_{i,max}$, higher rate connection i could be allowed to receive under the constrained optimization. The parameters $\lambda_{i,min}$ and $\lambda_{i,max}$ can be found through a lookup table that maps the QoS classes of connection i to the range of λ_i allowed. For ease of implementation, $\lambda_{i,min}$ and $\lambda_{i,max}$ can be chosen to be powers of 2, as used in the simulation below. The rate $R_{i,min}$ guaranteed for each connection can be read through the traffic descriptor of the connection.

As an illustrative example, we consider a scenario where there are five connections contending to get admitted into the system. $R_{a,max}$ is determined based on the congestion condition of the network, and is periodically updated. The connection characteristics and minimum admission requirements are shown in Table 1. The connection admission controller varies $\mu : 0 \leq \mu \leq 1$ and the rate shapers vary $\boldsymbol{\lambda} : \lambda_{i,min} \leq \lambda_i \leq \lambda_{i,max}$ to control the total system admission rate and each individual connection's admission rate through a geometric program.

Connection	$\lambda_{i,min}$	$\lambda_{i,max}$	$R_{i,min}$
1	0.21875	0.37500	0.15625
2	0.18750	0.31250	0.09375
3	0.25000	0.40625	0.15625
4	0.28125	0.43750	0.21875
5	0.09375	0.18750	0.06250

Table 1: Arrival traffic bounds and minimum rate requirements.

Figures 2 and 3 present simulation results illustrating how $(\boldsymbol{\lambda}, \mu)$ are dynamically optimized. In Figure 2, $(\boldsymbol{\lambda}, \mu)$ are chosen such that connection 1 admission rate is the largest possible while

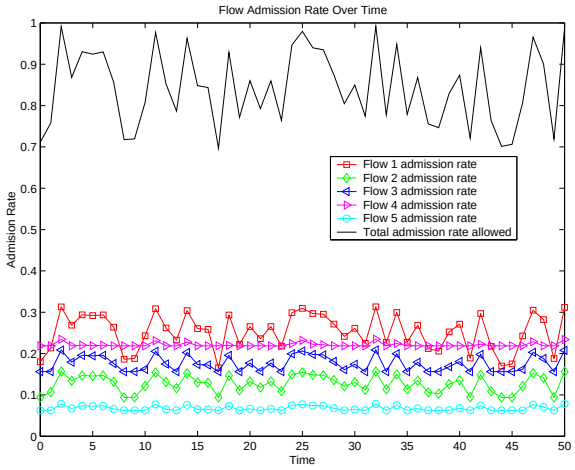

Figure 2: Optimized CAC: Maximize the admission rate for connection 1.

ensuring that the minimum admission rate requirements $R_{i,min}$ are met for all other connections, and that the total admission rate does not exceed the maximum rate Ra, max allowed by the system. Connection 1 is always favored over the other connections whenever possible under the QoS constraints. For instance, although connection 4 has a higher minimum admission rate requirement than connection 1, connection 1 is admitted more often.

In Figure 3, $(\boldsymbol{\lambda}, \mu)$ are chosen to maximize the minimum admission rate among all connections. With this objective, if there were no minimum admission rate requirements, all connections would have been admitted equally. However, because different connections have different characteristics and requirements, admission rates will vary. Connections that have relatively small minimum admission rate requirements (e.g., connections 2 and 5) are usually admitted at rates higher than requested. Intuitively, in Figure 3 all connections are treated as equally as possible, resulting in a narrower band of admission rate curves.

3.2 Application 2: Throughput optimization for wireless multihop networks

A more complicated example of the geometric programming method of efficient resource allocation is shown for power control in wireless networks where interference among the signals determine the QoS seen by the users. We focus on wireless multihop networks, where packets generally traverse several links from the source to the destination. Since the transmission environment can be different along each link, power control schemes must consider each link along a packet's path. The formulation used here explicitly takes into account the statistical variations of the received signal and the interference powers over a multihop network.

Consider a wireless multihop network with n transmitter/receiver pairs. Transmit powers are denoted as $P_1 \ldots, P_n$. Under Rayleigh fading, the power received from transmitter j at receiver i is given by $G_{ij}F_{ij}P_j$ where $G_{ij} \geq 0$ represents the path gain in the absence of fading. We also let G_{ij} encompass antenna gain and coding gain. The Rayleigh fading between transmitter j and receiver i is given by F_{ij}, which are assumed to be independent and have unit mean. The G_{ij}'s are appropriately scaled to accommodate variations from this assumption.

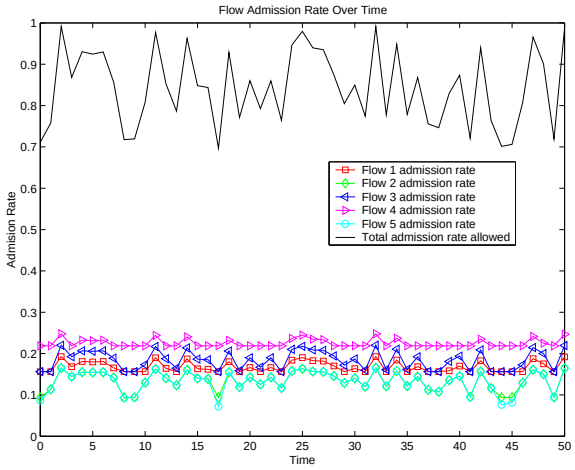

Figure 3: Optimized CAC: Maximize the minimum admission rate among all connections.

The distribution of the received power from transmitter j at receiver i is exponential with mean value $\mathbf{E}[G_{ij}F_{ij}P_j] = G_{ij}P_j$. The SIR for the receiver on link i is in the following GPA form:

$$\text{SIR}_i = \frac{P_i G_{ii} F_{ii}}{\sum_{j\neq i}^{N} P_j G_{ij} F_{ij} + n_i}. \tag{13}$$

We recall that the constellation size M used by a link can be closely approximated for MQAM modulation as follows: $M = 1 + \frac{-1.5}{\ln(5BER)}\text{SIR}$ where BER is the bit error rate. Defining $K = \frac{-1.5}{\ln(5BER)}$ leads to an expression of the data rate R_i on the ith link as a function of SIR: $R_i = \frac{1}{T}\log_2(1 + K\text{SIR}_i)$, which can be approximated in the high SIR regime as

$$R_i = \frac{1}{T}\log_2(K\text{SIR}_i). \tag{14}$$

The aggregate data rate for the system can then be written as the sum $\sum_i R_i = \frac{1}{T}\log_2\left[\prod_i K\text{SIR}_i\right]$. So in the high SIR regime, aggregate data rate maximization is equivalent to maximizing a product of SIR. This was also observed in [15] for optimizing throughput in cellular networks. The system throughput $R_{system} = \sum_i R_i$ is the aggregate data rate supportable by the system given a set of users with specified QoS requirements.

Outage probability is an important QoS parameter for reliability in wireless networks. A channel outage is declared and packets lost when the received SIR falls below a given threshold SIR_{th}, often computed from the BER requirement. Most systems are interference dominated and the thermal noise is relatively small, thus the ith link outage probability is

$$\begin{aligned} P_{o,i} &= \mathbf{Prob}\{\text{SIR}_i \leq \text{SIR}_{th}\} \\ &= \mathbf{Prob}\{G_{ii}F_{ii}P_i \leq \text{SIR}_{th}\sum_{k\neq i} G_{ik}F_{ik}P_k\}. \end{aligned}$$

The outage probability can be expressed as $P_{o,i} = 1 - \prod_{k\neq i}\frac{1}{1+\frac{\text{SIR}_{th}G_{ik}P_k}{G_{ii}P_i}}$ [9], which can be

approximated by

$$P_{o,i} = 1 - \prod_{k \neq i} \frac{G_{ii} P_i}{\mathrm{SIR}_{th} G_{ik} P_k} \tag{15}$$

when $\mathrm{SIR}_{th} G_{ik} P_k >> G_{ii} P_i$, *i.e.*, when there is no single dominant interferer.

Outage probability over a link also induces an outage probability over a path S:

$$P_{o,S} = 1 - \prod_{i \in S} (1 - P_{o,i}).$$

In wireless multihop networks with Rayleigh fading, we can use geometric programs to efficiently maximize system throughput under user throughput constraints and outage probability constraints.

Formulation 3 *The following nonlinear problem of optimizing power for system throughput maximization can be efficiently solved for global optimality as a geometric program:*

$$\begin{array}{ll} \textit{maximize} & R_{system}(\mathbf{P}) \\ \textit{subject to} & R_i(\mathbf{P}) \geq R_{i,min}, \ \ \forall i, \\ & P_{o,i}(\mathbf{P}) \leq P_{o,i,max}, \ \ \forall i, \\ & P_{o,S}(\mathbf{P}) \leq P_{o,S,max}, \ \ \forall S, \\ & P_i \leq P_{i,max}, \ \ \forall i \end{array} \tag{16}$$

where the optimization variables are the transmit powers $\mathbf{P}$.

The objective is to maximize the system throughput, which is equivalent to minimizing the posynomial $\prod_i \mathrm{ISR}_i$, where ISR is $\frac{1}{\mathrm{SIR}}$. The first constraint is from the data rate demand by each user. The second constraint represents the outage probability limitations demanded by users using single links. The third constraint represents the outage probability limitations for users using a multihop path. These inequality constraints put upper bounds on posynomials of $\mathbf{P}$, as can be readily verified through (14,15). The forth constraint is regulatory or system limitations on transmit powers. Thus (16) is indeed a geometric program, thus efficiently solvable for global optimality.

There are several obvious variations of Formulation 3 that maintain its geometric programming nature, *e.g.*, we can lower bound R_{system} as a constraint and maximize R_{i^*} for a particular user i^*, or maximize $\min_i R_i$ for maxmin fairness.

A simple four node multihop network, shown in Figure 4, is considered in the following numerical example. There are two connections $A \rightarrow B \rightarrow D$ and $A \rightarrow C \rightarrow D$. Nodes A and D, as well as B and C, are separated by a distance of 20m. Path gain between a transmitter and a receiver is the distance to the power -4. Each link has a maximum transmit power of 1W. All nodes use MQAM modulation. The baseband bandwidth for each link is 10kHz, the minimum data rate for each connection is 100bps, and the target BER is 10^{-3}. Assuming Rayleigh fading, we require outage probability be smaller than 0.1 on all links for an SIR threshold of 10dB. The CDMA spreading gain is 200. Using geometric programming, we find the maximized system throughput $R^* = 216.8$kbps, $R_i^* = 54.2$kbps for each link, $P_1^* = P_3^* = 0.709$W and $P_2^* = P_4^* = 1$W. The resulting SIR is 21.7dB on each link.

For this illustrative topology, we also consider a numerical example of admission control and pricing that were discussed in subsection 2.3. Initially the system has no users with QoS

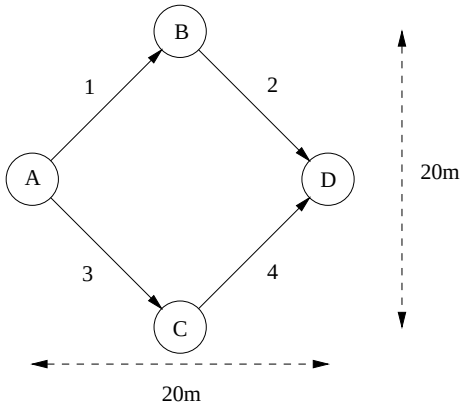

Figure 4: A small wireless multihop network.

constraints beyond the basic setup given previously. Three new users U_1, U_2, and U_3 are going to arrive to the system in order. U_1 and U_2 require 30kbps sent along the upper path $A \rightarrow B \rightarrow D$, while U_3 requires 10kbps sent from $A \rightarrow B$. All three users require the outage probability to be less than 0.1. When U_1 arrives at the system, the optimization with her QoS demands has the same solution as without the demands, so her price is the baseline price. Next, U_2 arrives, and her QoS demands decrease the maximum system throughput from 216.82kbps to 116.63kbps, so her price is the baseline price plus an amount proportional to the reduction in system throughput. Finally, U_3 arrives, and her QoS demands have no feasible solution, so she is not admitted to the system.

3.3 Application 3: Delay optimization for wireless multihop networks

The average delay a packet experiences traversing a network is another important design consideration in many applications. Queuing delay is often the primary source of delay, particularly for bursty data traffic in multihop networks, and is considered in this subsection to extend the scope of geometric-programming-based power control from the previous subsection.

A node i first buffers the received packets in a queue and then transmits these packets at a rate R set by the SIR on the egress link, which is in turn determined by the transmit powers $\mathbf{P}$. A FIFO queuing discipline is used here for simplicity. The approach can be extended to other disciplines. Routing is assumed to be fixed or only changes infrequently, and is feedforward with all packets visiting a node at most once. This restriction can be relaxed by recomputing the optimal transmit powers when routing changes.

Packet traffic entering the multihop network at the transmitter of link i is assumed to be Poisson with parameter λ_i and to have an exponentially distributed length with parameter Γ. Using the model of an $M/M/1$ queue, the probability of transmitter i having a backlog of $N_i = k$ packets to transmit is well known to be $\mathbf{Prob}\{N_i = k\} = (1-\rho)\rho^k$ where $\rho = \frac{\lambda_i}{\Gamma R_i(\mathbf{P})}$, and the expected delay is $\frac{1}{\Gamma R_i(\mathbf{P}) - \lambda_i}$. Under the feedforward routing and Poisson input assumptions, Burke's theorem [11] can be applied. Thus the total packet arrival rate at node i is $\Lambda_i = \sum_{j \in I} \lambda_j$

where I is the set of connections traversing this node. The expected delay $\bar{D}_i$ can be written as

$$\bar{D}_i = \frac{1}{\Gamma R_i(\mathbf{P}) - \Lambda_i}. \tag{17}$$

A bound $\bar{D}_{i,max}$ on $\bar{D}_i$ can thus be written as $\frac{1}{\frac{\Gamma}{T}\log_2(K\mathrm{SIR}_i)-\Lambda_i} \leq \bar{D}_{i,max}$, or equivalently, $\mathrm{ISR}_i(\mathbf{P}) \leq K2^{-\frac{T}{\Gamma}(\bar{D}_{max}^{-1}+\Lambda_i)}$, which is indeed an upper bound on a posynomial ISR of $\mathbf{P}$.

Formulation 4 *The following nonlinear problem of optimizing powers to maximize system throughput, subject to constraints on outage probability and expected delay, can be efficiently solved for global optimality as a geometric program:*

$$\begin{array}{ll} \textit{maximize} & R_{system}(\mathbf{P}) \\ \textit{subject to} & \bar{D}_i(\mathbf{P}) \leq \bar{D}_{i,max}, \ \ \forall i, \\ & \textit{Same constraints as in Formulation 3} \end{array} \tag{18}$$

where the optimization variables are the transmit powers $\mathbf{P}$.

The probability P_{BO} of dropping a packet due to buffer overflow at a node is also important in several applications. It is again a function of $\mathbf{P}$ and can be written as $P_{BO,i} = \mathbf{Prob}\{N_i > B\} = \rho^{B+1}$ where B is the buffer size and $\rho = \frac{\Lambda_i}{\Gamma R_i(\mathbf{P})}$. Setting a bound $P_{BO,i,max}$ on the buffer overflow probability also gives a posynomial constraint in $\mathbf{P}$: $\mathrm{ISR}_i(\mathbf{P}) \leq K2^{-\Psi}$ where $\Psi = \frac{T\Lambda_i}{\Gamma(P_{BO,i,max})^{\frac{1}{B+1}}}$.

Formulation 5 *The following nonlinear problem of optimizing powers to maximize system throughput, subject to constraints on outage probability, expected delay, and the probability of buffer overflow, can be efficiently solved for global optimality as a geometric program:*

$$\begin{array}{ll} \textit{maximize} & R_{system}(\mathbf{P}) \\ \textit{subject to} & P_{BO,i}(\mathbf{P}) \leq P_{BO,i,max}, \ \ \forall i, \\ & \textit{Same constraints as in Formulation 4} \end{array} \tag{19}$$

where the optimization variables are the transmit powers $\mathbf{P}$.

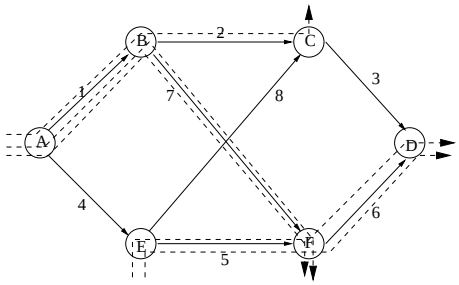

Figure 5: Network topology for delay constrained throughput maximization.

Consider the tradeoff between maximizing the system throughput and bounding the expected delay for the network shown in Figure 5. There are six nodes, eight links, and five

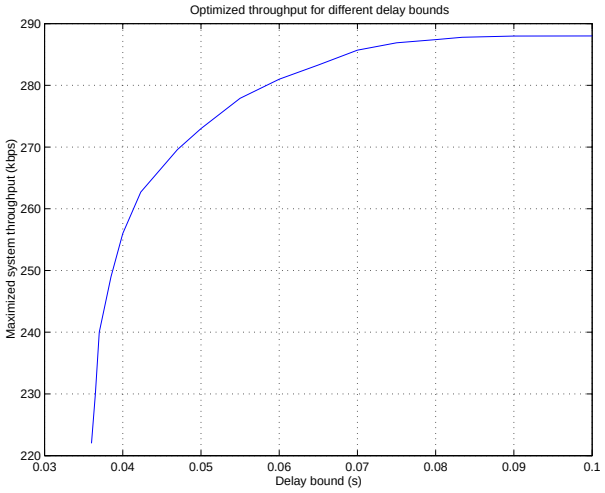

Figure 6: Optimal tradeoff between maximized system throughput and average delay constraint.

multihop connections. All sources are Poisson with intensity $\lambda_i = 200$ packets per second, and exponentially distributed packet lengths with an expectation of of 100 bits. The nodes use CDMA transmission scheme with a symbol rate of 10k symbols per second and the spreading gain is 200. Transmit powers are limited to 1mW and the target BER is 10^{-3}. The path loss matrix is calculated based on a power falloff of d^{-4} with the distance d, and a separation of 10m between any adjacent nodes.

Figure 6 shows the maximized system throughput for different upper bound numerical values in the expected delay constraints, obtained by solving a sequence of geometric programs, one for each point on the curve. There are no feasible power allocation to achieve delay smaller than 0.036ms. As the delay bound is relaxed, the maximized system throughput increases sharply first, then more slowly until the delay constraints are no longer active. Comparing performance with several existing power control algorithms (*e.g.*, the one in [7]), which cannot easily handle the nonlinear objective and constraints in Formulations 4 and 5, we find that either the delay bound is violated or the resulted throughput is not maximized by the existing algorithms. But geometric programming efficiently returns the globally optimal tradeoff between system throughput and queuing delay.

Obviously, the geometric programming method in this subsection can also efficiently computes the globally optimal power control if the objective is to minimize $\bar{D}_i$ or $P_{BO,i}$, subject to the constraints of lower bounds on system or individual throughput and upper bounds on per-link or per-path outage probability. All formulations in subsections 3.2 and 3.3 also directly apply to cellular wireless networks with only one-hop transmission from mobile users to the base station, extending the scope of power control problems solvable by the classic solution in CDMA systems that equalizes SIRs, and those by the iterative algorithm in [7] that minimizes total power subject to SIR constraints.

4 Conclusion

We show that a suite of resource allocations formulated as nonlinear, nonconvex optimization problems can be efficiently and globally solved through geometric programming and its conversion to convex form. Unlike general nonlinear problems that may take exponential amount of time to compute global optimality, these resource allocation problems can be solved in provably polynomial-time and often very fast in practice. This method can be applied to any allocation in the generalized proportional form, and extended to a variety of nonlinear functions of such forms. Specific formulations and numerical examples are provided to optimize a connection admission control scheme, and to determine the best allocation of powers in wireless multihop networks for different nonlinear objectives and under a variety of constraints, including system throughput, channel outage, queuing delay, and buffer overflow.

Acknowledgement

The author would like to thank Stephen Boyd, Arak Sutivong, Dan O'Neill, and David Julian at Stanford University for useful discussion and collaboration.

References

[1] S. Boyd and L. Vandenberghe, *Convex Optimization,* Cambridge University Press, 2003.

[2] M. Chiang and S. Boyd, 'Geometric programming duals of channel capacity and rate distortion,' To appear *IEEE Trans. Inform. Theory,* 2003.

[3] M. Chiang and A. Sutivong, "Efficient computation of fair resource allocation with QoS constraints," *Proc. IEEE Globecom,* San Francisco, CA, December 2003.

[4] M. Chiang, A. Sutivong, and S. Boyd, "Nonlinear optimizations of network queuing systems and connection admission control," *Proc. IEEE Globecom,* Taipei, ROC, November 2002.

[5] T. M. Cover and J. Thomas, *Elements of Information Theory,* Wiley, 1991.

[6] R. J. Duffin, E. L. Peterson, and C. Zener, *Geometric Programming: Theory and Applications,* Wiley, 1967.

[7] G. Foschini and Z. Miljanic, "A simple distributed autonomour power control algorithm and its convergence." *IEEE Trans. Vehicular Technology*, vol. 42, no. 4, 1993.

[8] D. Julian, M. Chiang, D. O'Neill, and S. Boyd, 'Resource allocation for constrained QoS provisioning in wireless cellular and ad hoc networks,' *Proc. IEEE Infocom,* New York, June 2002.

[9] S. Kandukuri and S. Boyd, "Optimal power control in interference limited fading wireless channels with outage probability specifications." *IEEE Trans. Wireless Comm.*, January 2002.

[10] F. P. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of Operations Research Society,* vol. 49, no. 3, pp.237-252, March 1998.

[11] L. Kleinrock, *Queueing Systems, vol. 1,* Wiley, 1972.

[12] S. H. Low, F. Paganini, and J. C. Doyle, "Internet congestion control," *IEEE Control Systems Magazine,* February 2002.

[13] Yu. Nesterov and A. Nemirovsky, *Interior Point Polynomial Algorithms in Convex Programming,* SIAM 1994.

[14] A. K. Parekh and R. Gallager, 'A Generalized processor sharing approach to flow control in integrated services networks: the single node case,' *IEEE Transactions on Networking,* vol. 1, no. 3, pp.344-357, June 1993.

[15] X. Qiu and K. Chawla, "On the performance of adaptive modulation in cellular systems." *IEEE Transaction on Communication*, pp. 884-895, June 1999.